A Program of Methodological Research to Arrive at the New International EQ-5D-5L Valuation Protocol

Oppe, Mark; Devlin, Nancy J.; van Hout, Ben; Krabbe, Paul F.M.; de Charro, Frank

Published in:
Value in Health

DOI:
10.1016/j.jval.2014.04.002

VALUE IN HEALTH 17 (2014) 445–453

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jval

METHODOLOGICAL ARTICLE

A Program of Methodological Research to Arrive at the New International EQ-5D-5L Valuation Protocol

Mark Oppe, PhD[1,5,*], *Nancy J. Devlin, PhD*[2], *Ben van Hout, PhD*[3], *Paul F.M. Krabbe, PhD*[4], *Frank de Charro, PhD*[5]

[1]*Institute for Medical Technology Assessment, Institute of Health Policy and Management, Erasmus University Rotterdam, Rotterdam, The Netherlands;* [2]*Office of Health Economics, London, UK;* [3]*HEDS, ScHARR, The University of Sheffield, Sheffield, UK;* [4]*Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands;* [5]*EuroQol Group Foundation, Rotterdam, The Netherlands*

ABSTRACT

Objectives: To describe the research that has been undertaken by the EuroQol Group to improve current methods for health state valuation, to summarize the results of an extensive international pilot program, and to outline the key elements of the five-level EuroQol five-dimensional (EQ-5D-5L) questionnaire valuation protocol, which is the culmination of that work. **Methods:** To improve on methods of health state valuation for the EQ-5D-5L questionnaire, we investigated the performance of different variants of time trade-off and discrete choice tasks in a multinational setting. We also investigated the effect of three modes of administration on health state valuation: group interviews, online self-completion, and face-to-face interviews. **Results:** The research program provided the basis for the EQ-5D-5L questionnaire valuation protocol. Two different types of tasks are included to derive preferences: a newly developed composite time trade-off task and a forced-choice paired comparisons discrete choice task. Furthermore, standardized blocked designs for the selection of the states to be valued by participants were created and implemented together with all other elements of the valuation protocol in a digital aid, the EuroQol Valuation Technology, which was developed in conjunction with the protocol. **Conclusions:** The EuroQol Group has developed a standard protocol, with accompanying digital aid and interviewer training materials, that can be used to create value sets for the EQ-5D-5L questionnaire. The use of a well-described, consistent protocol across all countries enhances the comparability of value sets between countries, and allows the exploration of the influence of cultural and other factors on health state values.

Keywords: EQ-5D, health-related-quality of life, quality-adjusted life-years, utility assessment.

Introduction

In the literature, the validity and reliability of the standard three-level EuroQol five-dimensional (EQ-5D-3L) questionnaire in many disease areas has been well documented (see, e.g., [1,2]). The EQ-5D-3L questionnaire has been shown to be a discriminative and evaluative measure in many conditions and has also been used in population health measurement. Some have argued, however, that generic measures such as the EQ-5D questionnaire may lack sensitivity or fail to capture important aspects of health in certain disease areas. These limitations could be attributed to the descriptive system, that is, the health dimensions and/or number of levels within each dimension, or the scoring algorithm. To address the issue of sensitivity, the EuroQol Group has undertaken an ambitious research and development program that aimed at the development of a more sensitive health status measurement instrument.

As a first step, the Group has developed the EQ-5D-5L questionnaire in which the number of levels in each dimension of the standard EQ-5D questionnaire is increased from three to five. The EQ-5D-5L questionnaire retains the original five dimensions but has modified the descriptive system to a five-level classification of severity: no problems, slight problems, moderate problems, severe problems, and extreme problems [3]. The EQ-5D-5L questionnaire describes 3125 (5^5) unique states compared with 243 (3^5) described by the EQ-5D-3L questionnaire.

Parallel fielding of the EQ-5D-5L questionnaire and the EQ-5D-3L questionnaire has provided a mapping algorithm between the two instruments, allowing EQ-5D-5L questionnaire states to be

* *Address correspondence to:* Mark Oppe, Institute for Medical Technology Assessment, Institute of Health Policy and Management, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR, Rotterdam, The Netherlands.
E-mail: oppe@euroqol.org.

1098-3015
http://dx.doi.org/10.1016/j.jval.2014.04.002

assigned values from existing EQ-5D-3L questionnaire value sets [4]. This provides an interim solution to value EQ-5D-5L questionnaire states. Of course, this solution has its limitations. Structurally, a value set based on preferences elicited specifically for the instrument is required. Value sets for the EQ-5D-3L questionnaire were mainly (although not exclusively) based on time trade-off (TTO) methods [5]. The TTO is, however, a complex valuation technique. The EuroQol Group has decided to explore the potential of a recently introduced valuation method in this field, namely, discrete choice (DC) modeling. The purpose is to explore whether this method can provide additional information to supplement the TTO values [6]. Furthermore, the conventional approach to TTO is known to have some important problems, particularly relating to the way values are obtained for health states considered to be worse than dead (i.e., values <0) [7,8]. For example, the conventional TTO uses conceptually different approaches to the valuation of states better than dead and worse than dead, resulting in arbitrarily large negative values. Traditionally, this has been redressed by a transformation of the negative values to a range with a minimum of -1 [7,8]. To address the issues, the EuroQol Group initiated a program of methodological research to develop new methods for TTO. Multiple variants of the "lead-time" and "lag-time" TTO [9–12] were investigated. The methodological research program also included experiments in the use of DC as a means of valuing EQ-5D questionnaire states [6].

On the basis of earlier methodological research, a draft protocol was developed to be tested and adapted after a multinational piloting program. A key objective of the pilot projects has been to test out variants of the interview in an international context and to adapt the methodology on the basis of the insights gained. The resulting protocol can be implemented in future EQ-5D-5L questionnaire valuation studies uniformly across different countries. A digital aid has been developed to implement the study design using the same valuation methods and protocol for data collection. This standardization will allow research teams engaging in valuation studies of the EQ-5D-5L questionnaire to produce results at a high level of quality control and comparability.

The aim of this article was to describe the international research conducted by the EuroQol Group to improve the methods for health state valuation, to summarize the results of an extensive international pilot program, and to outline the key elements of the EQ-5D-5L questionnaire valuation protocol, which is the culmination of that work. Because of the sheer volume of work presented in this article, we describe only the most important and relevant findings of the piloting program in this article. More detailed information on many aspects of the studies summarized here can be found in a forthcoming supplement [13].

Methods

Tasks

Time trade-off

The conventional TTO establishes the value for a health state by locating the amount of time in full health x, which is considered equal in utility terms to a given amount of time in a poor health state t, and calculating the value of the state as x/t. In EQ-5D questionnaire valuation studies, t is generally set (by convention) at 10 years. This approach has been widely used in valuation studies of the EQ-5D-3L questionnaire and works well to elicit values for states that are considered to be better than dead (i.e., have values between 0 and 1). The conventional TTO approach to eliciting values less than 0, however, is problematic. One means of avoiding the problems associated with the conventional approach to eliciting values less than 0 is to simply provide more "trading time" in full health so that when valuing very poor states of health, respondents can trade off more time. The additional trading time can be added either before the health state being evaluated (a "lead-time TTO") or after the state being evaluated (a "lag-time TTO") [9–12]. All TTO studies presented in this article used an iterative sequence to allow respondents to arrive at their points of indifference. The iteration sequence that was used in the main study is shown in Figure 1.

DC modeling

There has been renewed interest recently in ordinal response methods in estimating quality-of-life weights for various instruments [6,14–16]. Several empirical studies analyzing ordinal response data to infer latent cardinal values of the EQ-5D questionnaire have been reported in recent years [6,17–19]. Methods for collecting ordinal responses fall into the methodological tradition of DC analysis. The EuroQol Group started to explore the use of DC modeling as a valuation technique for the EQ-5D questionnaire in 2008 with a pilot study carried out in The Netherlands [6]. It was found that DC values broadly replicated the pattern found in TTO responses, although the DC values were consistently slightly higher than TTO values. The main difficulty in applying DC models was that these models generated values on an arbitrary scale, not on the metric of the quality (of life) component of the quality-adjusted life-year scale. This means that DC-based values need to be anchored on the utility scale, where full health has a value of 1 and dead has a value of 0. After these initial findings, it was decided to include a DC task in the pilot studies for the valuation of the EQ-5D-5L questionnaire to further investigate the merits of DC for health state valuation.

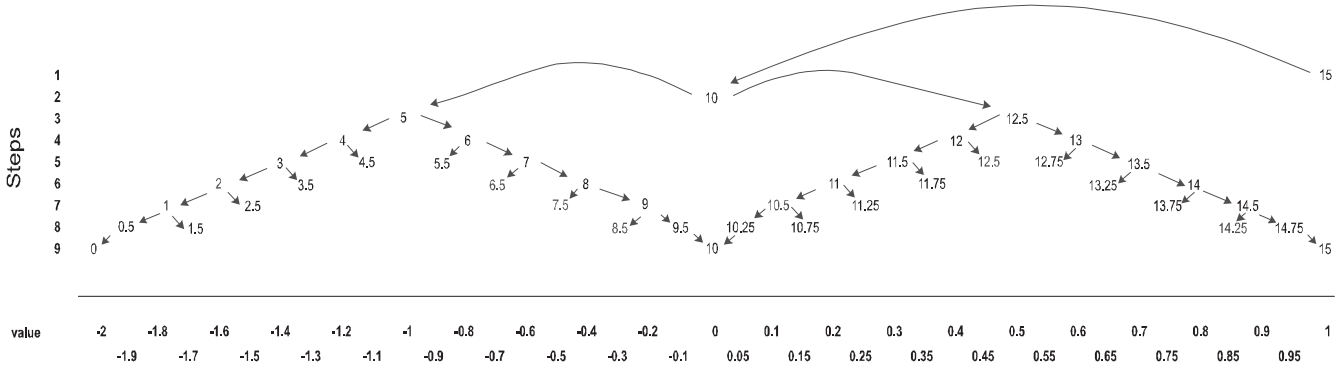

Fig. 1 – Iteration sequence for 10-year lead-time + 5-year disease-time TTO. Numbers represent the total time shown in the TTO task (lead time + disease time). TTO, time trade-off.

Pilot Studies

Multinational study (core)

The first pilot version of the valuation protocol for the EQ-5D-5L questionnaire was implemented in four countries: Canada, England, The Netherlands, and the United States (Table 1). This study consisted of a DC task with a design of 200 pairs (forced choice paired comparison of 2 EQ-5D-5L questionnaire states), a visual analogue scale (VAS) valuation task for the 400 states included in the 200 DC pairs, and a lead-time TTO task with 100 states (10-year lead-time duration and 5-year disease time duration). We used a blocked study design with 20 blocks so that each respondent would answer 10 DCs, 20 VASs, and 5 lead-time TTOs. The design for the DC and VAS was based on the Bayesian efficient design algorithm also used in the EQ-5D-3L questionnaire pilot study [6]. The design for the TTO task was created using Fedorov's exchange algorithm [20,21]. The sample size was 400 respondents for each of the four countries, totaling 1600 and leading to 80 observations per DC pair, per VAS state, and per TTO state. The experimental design was considered to result in adequate power. We used general population samples in all pilot studies. Given the pilot nature of the work, a formal sampling framework was not used, but respondents were selected in such a way that the study samples were broadly representative of the country populations with respect to age and sex (Table 1).

Multinational study (experimental elements)

The core study was extended to Argentina, China, Singapore, and Spain (Table 1). In the first three countries, experiments relating to the TTO task were included in the study; in the fourth country, an experiment with DC was included in the study. The initial choice for using the lead-time TTO as the prime TTO candidate was to a certain degree arbitrary. The EuroQol Group has more experience with lead-time TTO than with lag-time TTO. To make a more informed decision regarding this choice, the Argentina study contained an experiment to compare lead-time TTO and lag-time TTO. Also, the lead-time TTO approach might be susceptible to framing effects relating to the ratio of lead time to disease time [11]. Therefore, the Chinese study contained an experiment with 5-year lead-time and 5-year disease-time duration. Furthermore, the lead-time TTO might suffer from a framing effect relating to respondents' awareness of the distinction between the lead-time part (indicating the state is considered to be worse than dead) and the disease-time part (indicating the state is better than dead). To investigate this, the Singapore study contained an experiment in which the time window for the disease time and the lead time were separated more strongly in the visual display. All three studies used a split sample design in which half the respondents got the TTO task from the core study and half the respondents got the experimental TTO task. A subset of 10 states from the 100 state TTO designs of the core study was used in the TTO experiments. The 10 states were manually chosen so that they covered the entire utility range: 21111, 11221, 12112, 33133, 52221, 44113, 52324, 55523, 11145, 53555. The Spanish study included a DC experiment in which the paired comparisons between health states were complemented with comparisons between each health state included in the pairs and being dead. This allows the DC model to be anchored on the utility scale without the need of anchoring the DC model on TTO data [6,22]. However, the appropriateness of the model used for this analysis has been debated in the literature because respondents who prefer any state to death violate the assumptions underlying the DC framework [23]. An efficient subset of 50 pairs from the 200 pair design of the core study was generated with the Bayesian efficient design algorithm. The sample size was N = 400 for all four studies, resulting in 100 observations per study arm per state for the TTO experiments and 80 observations per pair for the DC experiment.

Follow-up study 1: Comparing TTO variants via an Internet panel

A follow-up study was undertaken in The Netherlands in the fall of 2011 to compare six different TTO tasks (Table 1). These were the conventional TTO, two variants of lag-time TTO (lag-time to disease-time ratios of 10 to 5 years and 10 to 10 years), and two variants of the lead-time TTO (lead-time to disease-time ratios of 10 to 5 years and 10 to 10 years). The sixth TTO task was also a lead-time TTO with a ratio of 10 to 5 years, but using an adapted iteration sequence to guide the respondents to their point of indifference. In addition to the TTO tasks, respondents got either a DC task or a case 2 best-worst scaling task (the latter is not presented in this article). The 100 state TTO design and 200 pair (400 state) DC design from the core study were used. The total sample size was N = 5000, or between 800 and 1000 respondents for each of the six TTO tasks. This implied between 40 and 50 observations per TTO state, which resulted in sufficient power for comparison of the six TTO tasks.

Follow-up study 2: Feasibility of composite TTO

A second follow-up study was undertaken in 2011 in The Netherlands to test the feasibility of the composite TTO approach (Table 1). This new type of TTO was implemented on the basis of our experiences from the previous pilot studies. In the composite TTO, values for states better than dead are elicited using the conventional TTO approach with a 10-year time frame, while values for states worse than dead are elicited using the lead-time TTO with a ratio of lead time to disease time of 1:1, that is, 10 years lead time and 10 years in the state to be valued. The 10 states selected to be valued in this study were the same as those in the TTO experiments of the multinational study. The sample size was N = 120, and each respondent valued all 10 states.

Data Collection Setting

In seven of the eight countries in the multinational study, data were collected in a group interview setting using respondents from marketing agencies. Respondents were asked to come to one of several alternative central locations. Instructions were then provided by trained interviewers to groups of 10 to 15 respondents, after which each respondent answered the questions in the interview. Interview assistance was available to answer respondents' questions and give support. In one country (England), the interview setting was changed to face-to-face interviews at the respondents' homes. For data collection in the first follow-up study, an existing online Internet panel was used. In the second follow-up study, to establish the feasibility of composite TTO, a face-to-face interviewer setting was used. The respondents for these interviews were asked by a marketing agency to come to a single central location where the interviews were conducted.

Valuation Technology

It is envisaged that the EQ-5D-5L questionnaire will be used in many different countries as is the case for the EQ-5D-3L questionnaire. Therefore, it is important that the valuation protocol can also be used internationally. The pilot valuation protocol was implemented in a digital setting from the start to make standardization of the protocol for the different experiments and different languages more feasible. The pilot studies were carried out by using the digital aid in which the pilot valuation protocol was embedded. The digital setting used a computer-assisted personal interview mode of administration: the EuroQol Valuation Technology (EQ-VT). The advantage of using a digital aid compared with more traditional pen-and-paper type "props" is that it forces interviewers to follow the same procedures, thereby

Table 1 – Overview of the EQ-5D-5L questionnaire valuation experiments.

	Study									
	Multinational study: Core				Multinational study: Experiments				Follow-up study 1	Follow-up study 2
Country	The Netherlands	England	United States	Canada	Argentina	China	Singapore	Spain	The Netherlands	The Netherlands
Data collection period	December 2010	January 2011	May 2011	May 2011	June 2011	June 2011	June 2011	June 2011	September 2011	November 2011
Setting	Group interview	1 on 1 interview	Group interview	Group interview	Group interview	Group interview	Group interview	Group interview	Internet survey	1 on 1 interview
Valuation tasks	DC, VAS, TTO	DC, VAS, TTO	DC, VAS, TTO	DC, VAS, TTO	DC, VAS, TTO	DC, VAS, TTO	DC, VAS, TTO	DC, TTO	BWS, DC, TTO	TTO
TTO variant	10-5 lead time	10-5 lead time	10-5 lead time	10-5 lead time	5-10 lag time 10-5 lead time	5-5 lead time 10-5 lead time	10-5 lead time	10-5 lead time	10-5 lead time 10-10 lead time 5-10 lag time 10-10 lag time Classic 10-y TTO 10-5 lead time different iteration procedure	Composite TTO: BTD classic 10-y TTO WTD 10-10 lead time
DC variant	Standard	Standard	Standard	Standard	Standard	Standard	Standard	DC + dead	Standard	–
Layout EQ-VT	Standard	Standard	Standard	Standard	Standard	Standard	Framing effect	Standard	Standard	Composite TTO
N states TTO	100	100	100	100	10	10	10	100	100	10
N pairs DC	200	200	200	200	200	200	200	50	200	–
N respondents	400	400	400	400	400	400	400	400	5000	120
% men	49	50	51	42	47	47	48	41	42	56
% age 18–35 y	30	57	45	48	47	53	47	35	33	36
% age 36–54 y	50	32	34	29	36	42	46	36	44	43
% age 55+ y	20	12	21	24	17	4	7	29	24	21

BWS, best-worst scaling; BTD, better than dead task; DC, discrete choice; TTO, time trade-off; VAS, visual analogue scale; WTD, worse than dead task.

reducing the probability of interviewer bias and eliminating data entry and coding errors. All elements of the protocol are implemented in the EQ-VT: assigning participants to sets of states from the underlying blocked design, randomization procedures, the iterative process in the TTO, and capturing and time stamping the participants' responses to all tasks. Different language versions of the EQ-VT were developed for which the same rigid guidelines for the translation of the EQ-5D questionnaire itself were used [24,25]. Last, the interviewer training materials were also standardized and officially translated.

Results

Multinational Study (Core)

Mean lead-time TTO values ranged from 0.73 to −0.23. The results of the multinational pilot study showed substantial clustering at value 0 and value 1 (and to a lesser extent also at −1, 0.5, 0.6, and 0.8) (Fig. 2). Of these, the clustering at 0 for mild and moderate states and at 1 for the severe states is counter-intuitive because this implies that many respondents value mild and moderate states as being as bad as dead and value severe states as being equal to full health. To get an impression of the response behavior of respondents, a cross-tabulation for all 100 TTO states included in the core part of the multinational study was constructed (Fig. 3). In this chart, the rows indicate the number of steps in the iteration sequence of the TTO taken by respondents to come to their final TTO answer. The columns indicate the corresponding TTO values for each iterative step (data from respondents using more than 20 iterative steps is excluded from this figure). As can be seen, 75% of the observations are contained within the 27 cells with more than 50 observations each, and a further 10% of the observations are contained in the 28 gray cells, containing between 20 and 50 observations each. This shows that most of the possible responses and iteration paths are rarely (or not at all) used by the respondents. In fact, 59% of all observations are distributed over only six values: $U = -1$, 0, 0.5, 0.6, 0.8, and 1 (6%, 19%, 10%, 6%, 6%, and 13%, respectively). Furthermore, many respondents gave their answer after only one or two steps in the iteration sequence (7% and 15%), indicating they "short-cut" the task. Comparing the face-to-face interview setting used in England with the group interview setting used in Canada, The Netherlands, and the United States, however, showed that the clustering was less prominent (i.e.. lower) for the face-to-face interviews:

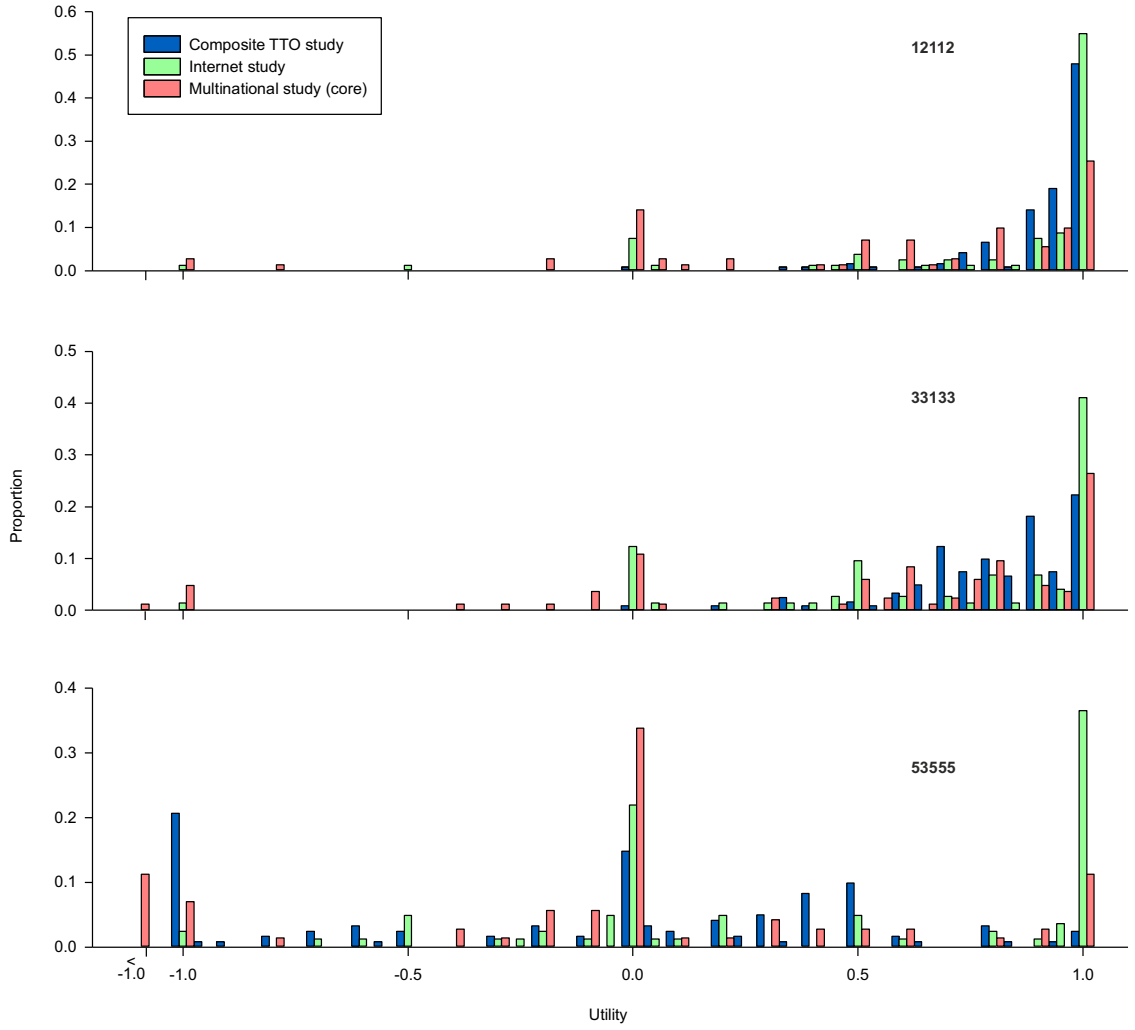

Fig. 2 – Comparison of the TTO values obtained in the multinational pilot study, the Internet study, and the composite TTO study. Observed proportion of responses for one mild state (12112), one moderate state (33133), and one severe state (53555) for the EQ-5D-5L questionnaire. EQ-5D-5L, five-level EuroQol five-dimensional.

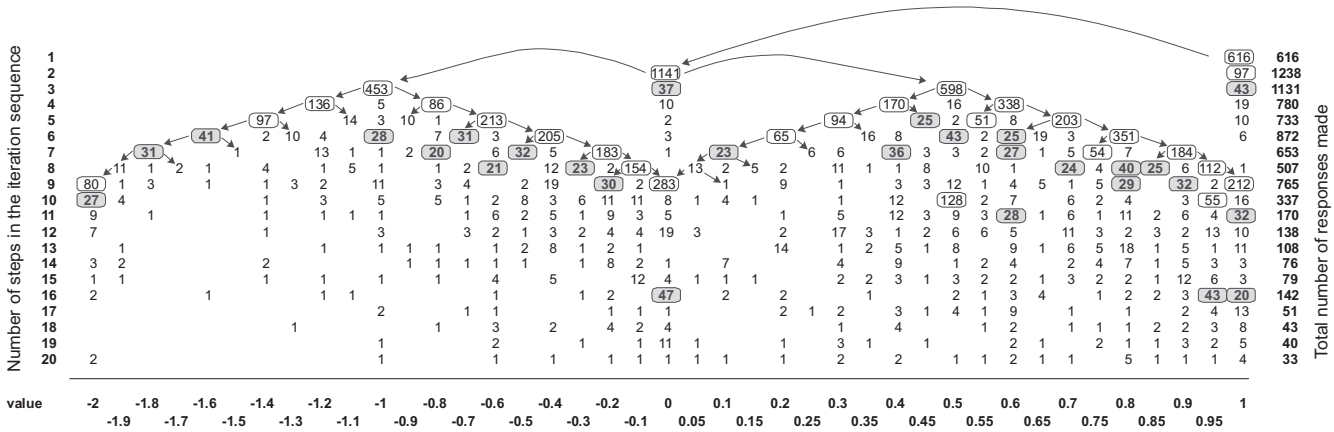

Fig. 3 – The number of steps in the iteration sequence made by the respondents (rows) versus the TTO-based utility value that was given as a final answer (columns). The figure shows the possible paths in the iteration sequence and how many responses were given for each of the paths. For example, the cell on row 2 and column 0 (i.e., steps = 2, utility = 0) indicates that in 1141 cases the point of indifference in the TTO task was reached at utility = 0 and that this point was obtained after two steps in the iteration sequence. The cell at (steps = 9, utility = 0) shows that 283 responses of utility = 0 were obtained after nine steps in the iteration sequence. Therefore, the same utility value was obtained but a different iteration path was used. Arrows indicate the possible steps in the indifference procedure, white cells contain more than 50 observations, and gray cells contain between 20 and 50 observations. TTO, time trade-off.

45% versus 63% for the six values and 8% versus 26% for the number of answers given after one or two steps in the iteration sequence. Because of the issues regarding the characteristics and/or quality of the TTO data, we could not reliably estimate and compare TTO models between countries.

To assess the performance of the DC task, we estimated and compared main-effects models in each country. For each country, we estimated a logit model with 20 dummy parameters (4 dummy variables for each of the five EQ-5D-5L questionnaire dimensions). All the 20 parameters were statistically significant, except level 2 for Usual Activities in The Netherlands ($P = 0.55$). Only three modest illogical ordering of regression coefficients were observed. Predictions for the complete set of 3125 EQ-5D-5L questionnaire states were quite similar for the four countries. Correlations between the countries were high: from 0.88 (The Netherlands vs. United States) through 0.97 (Canada vs. England). Some comparisons of countries showed higher dispersions, in particular between the values for The Netherlands and the United States. The comparability between Canada versus England and the United States was highest. Overall, the results from the DC task in the multinational pilot study showed that respondents were able to perform the task adequately in all countries as shown by the consistency of the resulting DC models between countries and attribute levels. Therefore, we decided to include the DC task in the final protocol.

Multinational Study (Experimental Elements)

Results from the Argentinean study showed that there were only slight differences in the mean observed values elicited using lead-time TTO and lag-time TTO and that the frequency distributions of the TTO responses were very similar, both suffering from similar levels of clustering as the core study (67% and 65%, respectively, for the six values compared with 59% from the core). These results did not provide enough evidence to establish superiority of either lead-time TTO or lag-time TTO. The Chinese study comparing lead-time TTO with a 10 to 5 ratio to that with a 5 to 5 ratio resulted in differences in the mean observed TTO values but not in the level of clustering of the data (60% and 62%, respectively, for the six values). In contrast, the split between the lead-time bar and disease-time bar in the framing effect study in Singapore resulted in a higher degree of clustering (70% and 75%, respectively, for the six values). Results from the Spanish DC experiment are not reported here, but can be found elsewhere [13].

Follow-Up Study 1: A Comparison of TTO Variants via an Internet Panel

In this study, data were collected via Internet without the presence of interviewers. This resulted in a worsening of the response clustering compared with the core study; in particular, the number of values = 1 was much higher (Fig. 2). Generally, lag-time TTO resulted in slightly lower values than did lead-time TTO, while extending the time frame from 15 years to 20 years resulted in slightly higher values for both lead-time TTO and lag-time TTO (results not shown). Shortcutting of the TTO task was more pronounced in the Internet study than in the core study: 26% of the respondents used one step in the iteration sequence ($U = 1$) and 20% used two steps ($U = 0$), compared with 7% and 15%, respectively, in the core study. Last, the adapted iteration procedure in the first follow-up study did not resolve the clustering of the data: clustering was still present, albeit at different values.

Follow-Up Study 2: The Feasibility of Composite TTO

The results from these studies prompted us to revise the protocol to address the data issues that were found. The results from the different pilot studies clearly indicated that it is not feasible to elicit TTO values in either a group interview setting or an online setting. Also, the lead-time approach showed that apart from theoretical benefits, it produced results that showed that a large proportion of respondents were using the complete time scale (i.e., lead time plus disease time) to trade off, even for states that evidently were not severe. This suggested that there would be merit in using the conventional approach to TTO to obtain values greater than 0, and introducing the lead-time TTO in which it becomes apparent that the value is less than 0. The second change was to deliberately opt for a face-to-face interview setting. Therefore, in the composite TTO study, conventional TTO was used to elicit values better than dead and lead-time TTO

to elicit values worse than dead. The choice for lead time over lag time was based on the fact that lead-time TTO is conceptually (and in practical terms from the participants' perspective) more in line with the conventional TTO task and on the fact that lag time was not shown to be superior to lead time. Because the time frame for the conventional TTO was set at 10 years, it was decided to use the lead-time TTO with a ratio of lead time to disease time of 10 to 10 years for the elicitation of values worse than dead. The results from the composite TTO study showed a marked improvement in data quality (Fig. 2). For the mild EQ-5D-5L questionnaire health state 12112, the clustering at $U = 0$ is no longer present and there are no values worse than dead. For the moderate state 33133, the clustering at $U = 0$ is almost completely gone while the clustering at $U = 1$ is markedly reduced. Again, there were no values given for this state that were worse than dead. For the severe state 53555, the clustering at $U = 1$ is almost completely absent, the clustering at $U = 0$ is reduced, but a new clustering appears at $U = -1$. This pattern of responses is in line with the expectations for such a severe health state.

EQ-5D-5L Questionnaire Valuation Protocol

Informed by the evidence from the multinational pilot studies, the EuroQol Group decided on a standardized protocol for EQ-5D-5L questionnaire value set studies. The protocol centers on systematic approaches to collecting values for EQ-5D-5L questionnaire health states. The first step in the valuation protocol consists of welcoming the respondent and explaining the purpose of the research he or she is taking part in (Table 2). Next, respondents are asked to complete the EQ-5D-5L questionnaire self-classifier, the EQ-VAS, and background questions regarding age, sex, and experience with illness. This introductory part is then followed by the first of the valuation tasks, the composite TTO (Fig. 4). After an explanation of how to interpret and carry out the composite TTO task, respondents are asked to evaluate 10 EQ-5D-5L questionnaire states, followed by three debriefing questions, for example, asking the respondents how difficult they found the composite TTO task. The respondents then receive instructions on how to carry out the DC task and are asked to complete seven paired comparisons (Fig. 5). This is then followed by three debriefing questions regarding the DC task. Last, the respondents are given an opportunity to comment on the entire valuation exercise and are thanked for their cooperation.

Table 2 – Elements of the EQ-5D-5L valuation protocol.

Start interview

1. General welcome
2. Introduction
   a. Self reported health on the EQ-5D-5L descriptive system
   b. Self reported health on the EQ-VAS
   c. Background questions
3. Composite Time Trade-Off
   a. Instructions and example of TTO task
   b. TTO valuation of 10 EQ-5D-5L states
   c. TTO debriefing/structured feedback
4. Discrete Choice
   a. Instructions and example of DC task
   b. DC valuation of 10 pairs of EQ-5D-5L states
   c. DC debriefing/structured feedback
5. General thank you and goodbye

End interview

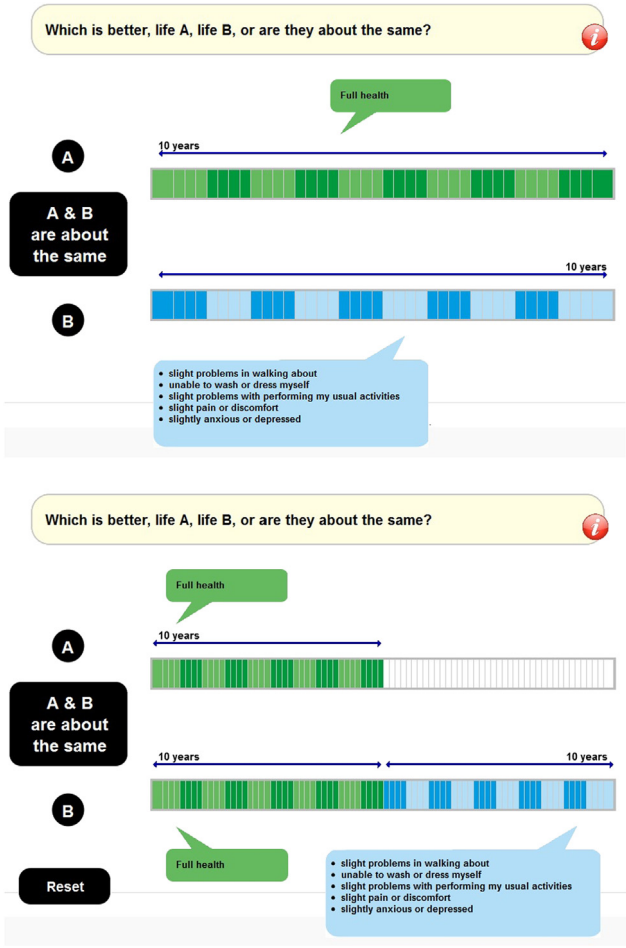

Fig. 4 – The composite TTO task. Conventional TTO with a 10-year time frame is used to value states better than dead, and lead-time TTO with a time frame of 20 years is used to value states worse than dead. TTO, time trade-off.

Discussion

The protocol described in this article represents the culmination of an ambitious program of research commissioned and coordinated by the EuroQol Group over a 3-year period. This has enabled us to make considerable progress in improving the methods used to obtain TTO values; in complementing those by using additional information on preferences obtained by DCs; and in developing explicit study designs to underpin the selection of states and tasks.

For the first time, the EuroQol Group has developed a standard protocol embedded in a digital aid and accompanied by interviewer training materials, which can be made available to study teams in countries wishing to develop local value sets for the EQ-5D-5L questionnaire. Valuation studies have already been undertaken for Canada, China, England, The Netherlands, and Spain. More studies are in the process of development. The use of a well-described, consistent protocol across all countries will ultimately create a unique opportunity to compare health state

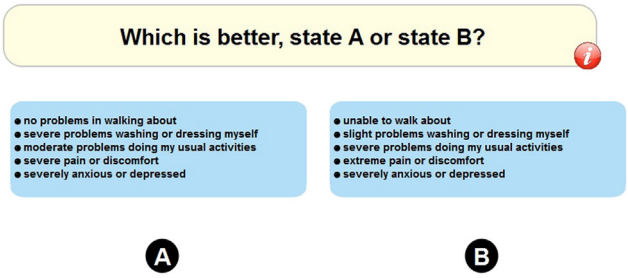

Fig. 5 – The discrete choice task, a forced choice paired comparison of two EQ-5D-5L questionnaire health states. EQ-5D-5L, five-level EuroQol five-dimensional.

preferences between countries and to explore the influence of cultural and other differences on health state values.

A key methodological issue considered in the development of this protocol was whether computer-based technology designed to present the valuation tasks could replace the requirement for face-to-face interviews by facilitating self-completion online or in group-based settings. This would have the considerable advantage of reducing costs and time involved in conducting face-to-face interviews. Our research, however, clearly indicated the importance of trained expert interviewers during TTO valuation tasks, and this is likely to remain indispensable for TTO tasks that involve a process of iteration toward indifference. Our results correspond to those of Norman et al. [26] who also observed clustering of values around $U = 1$, $U = 0$, and $U = -1$ in an online TTO valuation study.

There is no "perfect" method for eliciting stated preferences for health states: all methods have attendant advantages and disadvantages, and inevitably choices between methods must reflect judgments about these relative merits, given available evidence. Inevitably, there remain methodological questions that can be addressed in future research that might further improve the valuation protocol. The composite TTO has improved the means of eliciting values worse than dead and has removed the need for arbitrary rescaling of values required by the conventional TTO. This was accomplished by using the linear scale restricted to −1 of the lead-time TTO, with a ratio of lead time to disease time of 10 to 10 years. Research on the lead-time TTO using different lead-time to disease-time ratios (reflected by different lower bounds for the resulting utility scale) showed, however, that the values worse than dead are affected by the choice of this ratio. In other words, some participants considering very severe states of health might want to trade off more time than the maximum possible within the design of the composite TTO task. Further research will be undertaken to explore methods that might be used to model these "censored" values [11].

Unresolved issues also remain for DC methods, such as whether respondents use decision heuristics thereby violating the DC model assumptions and the fact that for linear-in-parameter models, the scaling factor φ and the parameter estimates β cannot be estimated separately, but are estimated as a ratio $\gamma = \beta/\varphi$ [27,28]. This means that β and φ are perfectly confounded, which can bias the parameter estimates. This issue is also known as *variance heterogeneity* [23]. Alternative formulations of DC methods could improve the application of DC methods for the quantification of health states. These include simply stating the duration that applies to all states; including duration as an attribute to be varied within the design [23,29]; asking participants to state whether the states in the paired comparisons are worse than being dead; or to identify the aspect of each health state that is best and worst (i.e., case 2 best-worst scaling) [30]. These approaches can potentially be used to provide quantitative estimates of the utility associated with different dimensions and levels of the EQ-5D-5L questionnaire and to overcome the issue in DC of anchoring values at 0 and 1. Experimentation in this area is continuing and may provide a basis for future approaches to valuation.

There is an ongoing scientific debate on whether the scales on which the elicited utilities are placed are true interval scales or not, as is required by the quality-adjusted life-year model. In the case of TTO, this is made apparent by the discussion on constant proportional trade-offs [31,32]. It equally applies, however, to other valuation techniques, including standard gamble and DC.

For the first time, the EuroQol Group has developed a standard protocol with accompanying digital aid and interviewer training materials to elicit values for the EQ-5D-5L questionnaire. The use of a well described, consistent protocol across all countries enhances the comparability of value sets between countries, and allows the exploration of influence of cultural and other differences on health state values. Researchers interested in undertaking a valuation study for the EQ-5D-5L questionnaire using the EQ-VT can contact the EuroQol Group (www.euroqol.org) for further information on obtaining and using the EQ-VT.

Acknowledgments

We thank all the team members for their participation in the studies: Argentina: Federico Augustovski, Lucila Rey Ares, and Vilma Irazola; Canada: Feng Xie and Kathryn Gaebel; China: Minghui Li and Gordon Liu; England: Koonal Shah, Andrew Lloyd, and Paul Swinburn; The Netherlands: Bas Janssen, Elly Stolk, and Matthijs Versteegh; Singapore: Nan Luo; Spain: Juan Manuel Ramos Goni; United States: Simon Pickard; EuroQol support and IT team: Gerben Bakker, Job de Bruyne, and Arnd Jan Prause.

Source of financial support: The study in England reported in this article was funded by the Department of Health (Policy Research Programme grant PRP 070-0065). Follow-up study 1, the Internet-based TTO study in The Netherlands, was funded by ZonMW. The other studies were funded by the EuroQol Group. Views expressed in the article are those of the authors and the EuroQol Group, and not of the Department of Health or ZonMW.

REFERENCES

[1] Pickard AS, Wilke CT, Lin HW, Lloyd A. Health utilities using the EQ-5D in studies of cancer. Pharmacoeconomics 2007;25:365–84.
[2] Janssen MF, Lubetkin EI, Sekhobo JP, Pickard AS. The use of the EQ-5D preference-based health status measure in adults with type 2 diabetes mellitus. Diabet Med 2011;28:395–413.
[3] Herdman M, Gudex C, Lloyd A, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011;20:1727–36.
[4] van Hout B, Janssen MF, Feng YS, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. Value Health 2012;15:708–15.
[5] Szende A, Oppe M, Devlin N, ed. EQ-5D values sets. Inventory, Comparative Review and User Guide. Dordrecht: Springer, 2007.
[6] Stolk EA, Oppe M, Scalone L, Krabbe PF. Discrete choice modeling for the quantification of health states: the case of the EQ-5D. Value Health 2010;13:1005–13.
[7] Lamers LM. The transformation of utilities for health states worse than death: consequences for the estimation of EQ-5D value sets. Med Care 2007;45:238–44.
[8] Tilling C, Devlin N, Tsuchiya A, Buckingham K. Protocols for time tradeoff valuations of health states worse than dead: a literature review. Med Decis Making 2010;30:610–9.
[9] Robinson A, Spencer A. Exploring challenges to TTO utilities: valuing states worse than dead. Health Econ 2006;15:393–402.
[10] Devlin NJ, Tsuchiya A, Buckingham K, Tilling C. A uniform time trade off method for states better and worse than dead: feasibility study of the 'lead time' approach. Health Econ 2011;20:348–61.

[11] Devlin N, Buckingham K, Shah K, et al. A comparison of alternative variants of the lead and lag time TTO. Health Econ 2013;22(5):517–32.
[12] Attema AE, Versteegh MM, Oppe M, et al. Lead time TTO: leading to better health state valuations? Health Econ 2013;22:376–92.
[13] Eur J Health Economics 2013;14(Suppl. 1): page range.
[14] de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Econ 2012;21:145–72.
[15] Coast J, Flynn TN, Natarajan L, et al. Valuing the ICECAP capability index for older people. Soc Sci Med 2008;67:874–82.
[16] Netten A, Burge P, Malley J, et al. Outcomes of social care for adults: developing a preference-weighted measure. Health Technol Assess 2012;16:1–166.
[17] Hakim Z, Pathak DS. Modelling the EuroQol data: a comparison of discrete choice conjoint and conditional preference modelling. Health Econ 1999;8:103–16.
[18] Salomon JA. Reconsidering the use of rankings in the valuation of health states: a model for estimating cardinal values from ordinal data. Popul Health Metr 2003;1(1):1–12.
[19] Krabbe PFM. Thurstone scaling as a measurement method to quantify subjective health outcomes. Med Care 2008;46:357–65.
[20] Fedorov V. The design of experiments in multiresponse case. Theory Probab Appl 1971;16:323–32.
[21] Fedorov V. Theory of Optimal Experiments. New York: Academic Press, 1972.
[22] McCabe C, Brazier J, Gilks P, et al. Using rank data to estimate health state utility models. J Health Econ 2006;25:418–31.
[23] Flynn TN. Using conjoint analysis and choice experiments to estimate QALY values: issues to consider. Pharmacoeconomics 2010;28:711–22.
[24] Herdman M, Fox-Rushby J, Rabin R, et al. Producing other language versions of the EQ-5D. In: Brooks R, Rabin R, de Charro F, eds., The Measurement and Valuation of Health Status Using EQ-5D: A European Perspective. Kluwer Academic Publishers, 2003.
[25] Rabin R, Herdman M, Fox-Rushby J, Badia X. Exploring the results of translating the EQ-5D into 11 European languages. In: Brooks R, Rabin R, de Charro F, eds., The Measurement and Valuation of Health Status Using EQ-5D: A European Perspective. Kluwer Academic Publishers, 2003.
[26] Norman R, King MT, Clarke D, et al. Does mode of administration matter? Comparison of online and face-to-face administration of a time trade-off task. Qual Life Res 2010;19:499–508.
[27] Mazzotta MJ, Opaluch JJ. Decision making when choices are complex: a test of Heiner's hypothesis. Land Econ 1995;71:500–15.
[28] Yatchew A, Griliches Z. Specification error in probit models. Rev Econ Stat 1985;67:134–9.
[29] Bansback N, Brazier J, Tsuchiya A, Anis A. Using a discrete choice experiment to estimate health state utility values. J Health Econ 2012;31:306–18.
[30] Flynn TN, Louviere JJ, Peters TJ, Coast J. Best-worst scaling: what it can do for health care research and how to do it. J Health Econ 2007;26:171–89.
[31] Attema AE, Brouwer WB. On the (not so) constant proportional trade-off in TTO. Qual Life Res 2010;19:489–97.
[32] Attema AE, Brouwer WB. Constantly proving the opposite? A test of CPTO using a broad time horizon and correcting for discounting. Qual Life Res 2012;21:25–34.